# **Strategic Thinking**

# **Semester Two**

# **Capstone Project Report**

**Marie Pigott sbs23093**

**Audrey Fayne mda21514**

(6,000 words)

**2023**

**CCT College**

**Dublin, Ireland**

## 

## **Table of Contents:**

1. Introduction
2. Business Understanding and Goals
3. Data Understanding
4. EDA and Data Preparation
5. Models and Machine Learning Algorithms

1. Results and Model Evaluation
2. Conclusion
3. References

Annex A : Data Dictionary

## **Table of Figures**

Figure 1. CRISP-DM Process Model of Data Mining (from Martínez-Plumed et al, 2021)

Figure 2. BMI data from Healthy Ireland Summary Report 2019.

Figure 3. EDA Box Plots.

Figure 4. Modelling Research.

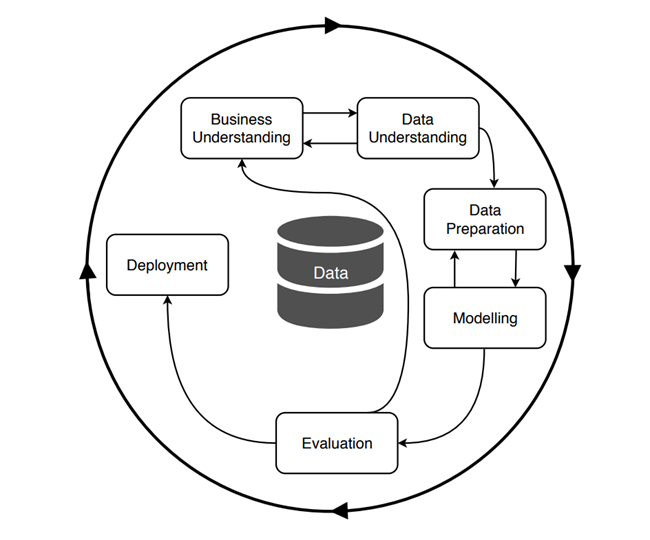
Figure 5. Data Information.

Figure 6. Team Contributions by member.

### 

### **Introduction**

* 1. The following is a report written to explain and describe the work completed for the Strategic Thinking Capstone Project. It includes work completed in semester one which has been enhanced and added to to improve the quality of the results.
  2. To structure our approach to this continuous assessment project we followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) Methodology. The headings in the subsequent report follow the phases of the CRISP-DM Method. We omitted the deployment phase as it is not a real-life business case. It was replaced by a reflection and learning phase.



*Figure 1. CRISP-DM Process Model of Data Mining (from Martínez-Plumed et al, 2021)*

### **Business Understanding**

The business area we have chosen to conduct data analysis on is health. The specific area we looked at was diabetes due to having some previous subject matter understanding.

The business problem we identified was the need to know how best to predict the development of type 2 diabetes.

This business problem can be translated into the data analytics problem which was ‘Can we create a predictive model to indicate if a person will develop diabetes based on potential risk factors such as obesity, high cholesterol and high triglyceride?’

* 1. **Background**
     1. Ireland has one of the highest rates of obesity in the EU, as measured by BMI (body mass index). According to data from 2019, 25.7% of Irish females and 26% of Irish males were obese when compared to the EU averages of 16% and 18% respectively (Eurostat, 2021). Data from the Healthy Ireland Survey analysis also for 2019 indicated an overall obesity rate of 23% in the surveyed population as illustrated in Figure 2 below.

Chart

Description automatically generated

*Figure 2. BMI data from Healthy Ireland Summary Report 2019*

* 1. The factors which lead to the development of diabetes are likely multifactorial but obesity in adults is considered a risk factor for premature mortality from diabetes and also heart disease and cancer (Blüher, 2019). The direct costs of treating these diseases, along with the associated costs such as decreased productivity have been estimated at 1.13 billion or 2.7% of the total health expenditure in Ireland in 2011 (Safefood, 2012). These costs, plus the impact on people’s lives and health are enormous and hence it would be useful to identify risk factors and work to prevent these conditions from occurring. Machine learning has been used recently in the areas of diabetes research (Woldaregay et al, 2019; Zevi et al., 2015) and has great potential for prediction in this area.
  2. **Project Goal**
     1. Our goal is to create a predictive machine learning model to indicate if a person will develop diabetes based on potential risk factors such as obesity, high cholesterol and high triglyceride?’use Machine Learning (ML) to identify if obesity is a strong determinant of diabetes.
     2. The dataset we have chosen is a Diabetes Dataset which has information ono whether a patient has diabetes, including clinical diagnostic measurements. The dataset includes information on whether the subjects have diabetes and also information on BMI. It is covered by CC BY 4.0 licence description and is located at <https://data.mendeley.com/datasets/wj9rwkp9c2/1>
     3. The data attributes are outlined in the data dictionary and include medical information and laboratory analysis such as No. of Patient, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Patient's diabetes disease class.
  3. **Hypothesis**
     1. Can we create a predictive model which can tell if a person will develop diabetes based on potential risk factors such as obesity, high cholesterol and high triglycerides
     2. Target Variable: CLASS
     3. Predictors: age, gender, BMI, Chol, TG etc.
     4. CLASS = N No diabetes CLASS = D Diabetes

We need to look at distribution of target variable

**Can we predict:**

**H0:** A person will develop diabetes due to specified risk factors.

**H1:** A person will not develop diabetes due to specified risk factors.

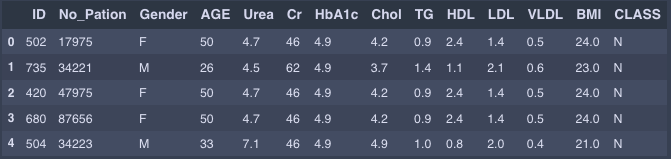
* 1. **Our approach**

We will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, going through the stages of Business Understanding (this document), Data Understanding (initial data shows various diagnostic measurements, no missing data but some values are 0 which does not make sense from a ‘business’ view point (i.e. blood pressure cannot be 0), Data Preparation (we will clean the data and analyse it using various libraries including: Panda, Numpy, Seaborn, and Matplotlib), Modelling and Evaluation and Deployment - We aim to use machine learning models and algorithms to find which one shows the best prediction and most accurate results. We intend to start with linear regression, and also the random forest algorithm because they are useful for regression and classification problems.

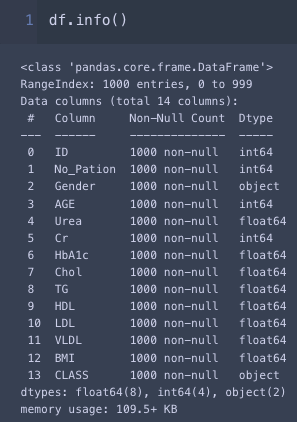
* 1. **Success criteria**
     1. Measures of success of this analysis project will include:
        1. Successfully cleaning the Data to make it possible to use in machine learning models.
        2. Successfully using data visualisations to explore the data.
        3. Successfully implementing machine learning models. We will be successful if we produce a model with high accuracy; neither over nor under fitting, and which can be applied to another dataset with even more features.
  2. **Sprint Sessions**
     1. There was a sprint Session with James Garza on the 13th April to discuss our continuous assessment project. Marie explained our idea of using one of two datasets based on diabetes information. The first was the ‘Pima’ dataset based on a study of Pima Indians and the other was called ‘Dataset of Diabetes. Csv’ based on a study from Iraq. or the other Dataset to look at diabetes and the correlation with obesity. James advised us that the PIMA dataset has been overused and that there would be little new information to gain from analysis. He recommended that we go with the Diabetes dataset based on a study in Iraq and that we put further thought into the complexity of the problem.

### **Data Understanding**

* 1. **Collecting initial data and loading into library**
     1. The ‘Diabetes Dataset .csv’ was downloaded from Mendeley Data and uploaded into the Juypter notebook.



* 1. **Describe the Data**
     1. The source of the data was called ‘Diabetes Dataset .csv’. (Rashid, Ahlam, 2020),
     2. Using “.shape” and “.info” functions we were able to get information on the dataset as follows:
        1. Number of Attributes: 14
           1. Numerical: 12
           2. Float: 8
           3. Object: 2
        2. Number of observations: 1,000.

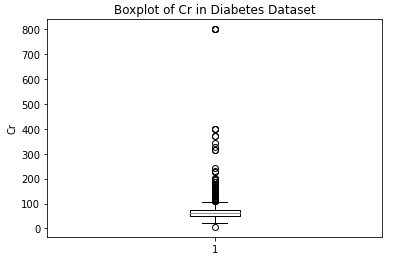
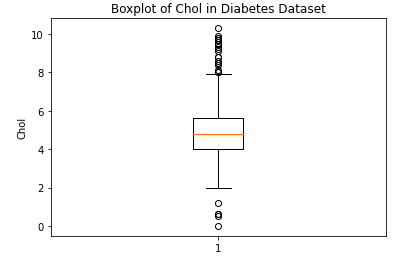
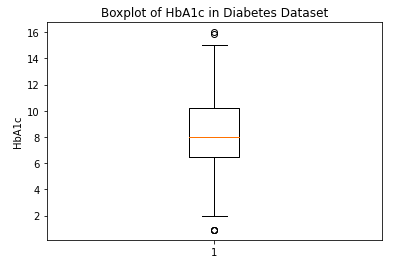
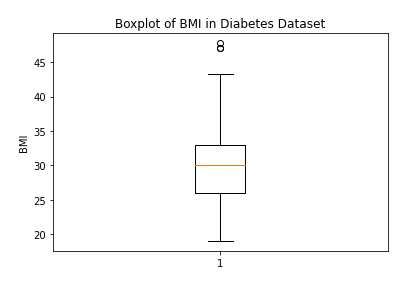
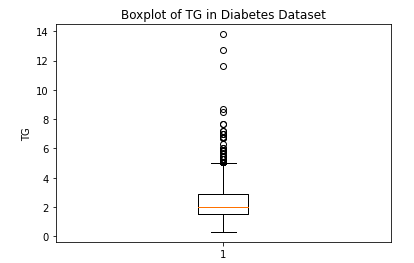
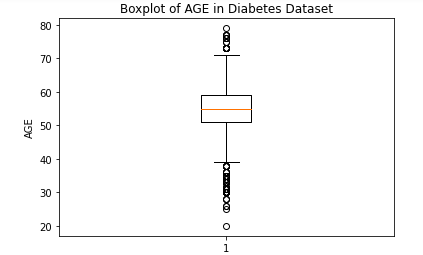


We created a data dictionary to help us understand the data so we could plan how to treat it. For example, we identified what the normal range of a test would be to help us determine whether we should treat that result as an outlier or keep it. Ultimately we kept all of the biological data other than values of 0 which were biologically impossible, and dropped only the ID numbers as being unique numbers they did not add to the accuracy of the machine learning models applied.

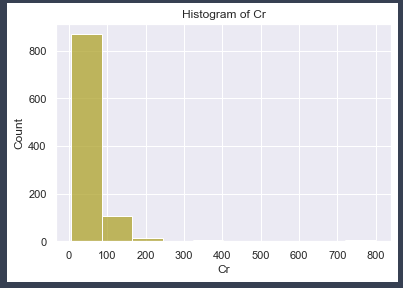
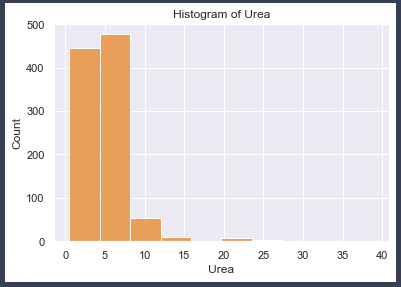
* 1. **Explore the Data**
     1. **We used the following libraries in Python:** Pandas, NumPy, Matplotlib, Seborn, Sklearn.
     2. We checked for duplicates - there were none.
     3. We checked for 0 values.

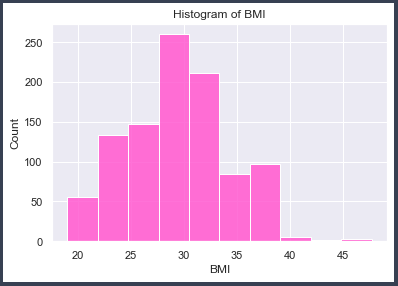
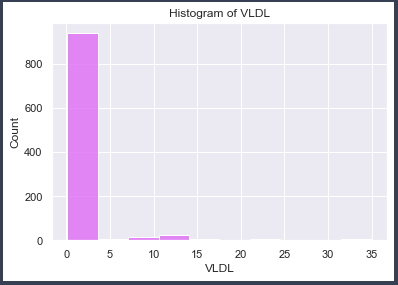
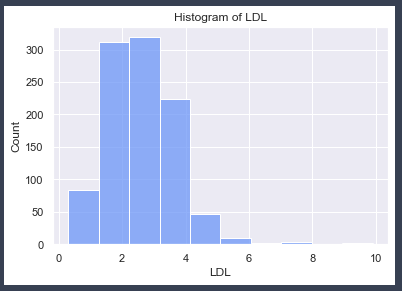
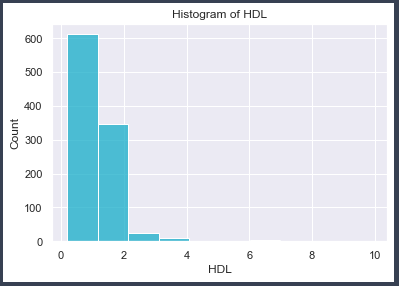
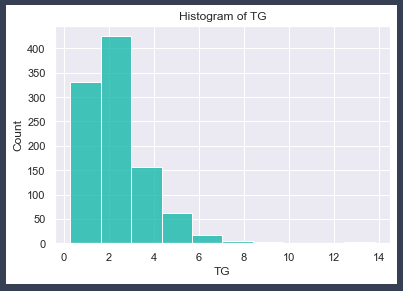
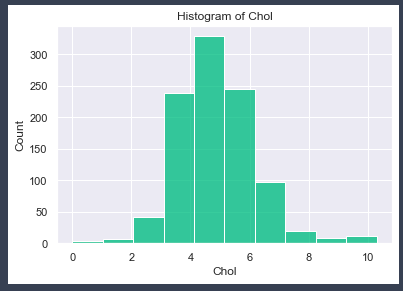
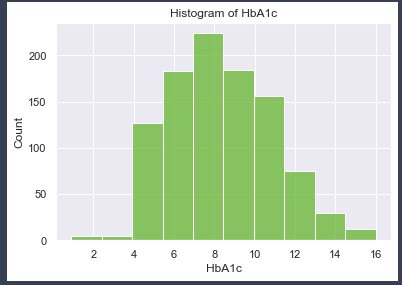
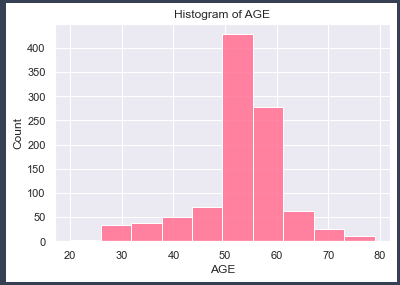
### **Data Preparation**

* 1. EDA visualizations: The first EDA visualisation we did was to make box plots of all of the attributes with numeric values. This was done to see how the data was distributed and especially to see if there were significant outliers.



*Figure 3. EDA Boxplots*





*Figure. 4 Histograms*

* 1. Data Cleaning
     1. We removed the two unique identifier numbers because they are not relevant data for the machine learning models. There was a patient number and an ID.
  2. Format data
     1. Convert string values to numeric values by using one hot encoder and dummies. For the Gener category this divided the attributed into three - Gender\_M, Gender\_F and Gender\_f. This showed that there were inconsistencies in the way female was entered in the dataset. This cause a lot of confusion and errors in the code.
     2. The column ‘Class’ which is the target variable had three options of Y, N or P. Dummies was also used on this.

### 

### **Modelling**

* 1. **Modelling Technique Research**

Why we chose what we chose : Previous work shows commonly used ML algorithms Naïve Bayes and Random Forest seemed to demonstrate good accuracy of Diabetes as shown by the following:

| **Researcher** | **Model** | **Database used** |
| --- | --- | --- |
| Soni & Varma, 2020 | Predicted diabetes using SVM, KNN, Decision tree, Random forest, logistic regression, and gradient boosting with 77% classification accuracy | Pima |
| Singh et al, 2017 | Predicted diabetes using a correlation-based feature selection technique to remove irrelevant features and then used a function-based MLP, probabilistic-based Naïve Bayes, decision tree-based random forest. They got 79.69% prediction accuracy for Naïve Bayes. | Pima |
| Joshi and Chawan (2018) | Implemented an early diabetes prediction model using logistic regression, SVM, and ANN networks. | UCI diabetes dataset |
| Sisodia and Sisodia (2018) | Naïve Bayes, SVM, and decision tree. Naïve Bayes outperformed with an accuracy of 76.30% | Pima |
| Sarathambekari et al (2022) | Logistic Regression, Gaussian NB, KNN, SVM, Decision Tree, Random Forest and Gradient Boosting classification algorithms. Random Forest provided accuracy of 98% | Comparative analysis - Mostly PIMA |

*Figure 4. Modelling Research*

* 1. When looking specifically at obesity as a predictor, the following were found:

| **Researcher** | **Model** | **Database used** |
| --- | --- | --- |
| Tan et al., 2022 | LASSO method used to screen for risk variables, and to construct a nomogram incorporating the selected risk factors in the training group | DATADRYAD database |
| Cai et al., 2021 | LASSO method used to screen for risk variables and construct a model | No of databases |
|  | Used Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes to identify the presence of obesity using publicly available health data, using a novel approach with sophisticated ML methods to predict obesity as an attempt to go beyond traditional prediction models, and compared the performance of three different methods.  Logistic Regression method shows the highest performance | No of databases |

* 1. **Modelling Technique Research**
     1. We realised that we needed to create a supervised ML classification model, as the target variable is categorical
     2. We used a multi-class classification model if we use an appropriate algorithm such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), Naive Bayes, and Neural Networks.
     3. The data has 103 (no-diabetes), 53 (pre-diabetic) and 844 (diabetic) patients. Pre-diabetic is a state in which some blood results show poor glucose control but the patient has not been diagnosed with diabetes. So for the purposes of our work we can consider pre-diabetic as no-diabetes because that is technically true.

| **FEATURE NUMBER** | **COlUMN HEADING** | **ATTRIBUTE NAME** | **ATTRIBUTE TYPE** | **Measurement value** | **Reported Range** | **Normal Range** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Gender | Gender | Character |  | M and F | M and F |
| 2 | Age | Age | Integer | Years | 20 – 79 | Up to 100 |
| 3 | Urea | Urea | Numeric | mg/dl | 0.5 - 38.9 | 2.5 – 8.5 |
| 4 | Cr | Creatinine | Integer | micromol/l | 48 – 80 | 49-90 |
| 5 | HbA1c | Haemoglobin A1c | Numeric | mmol/l | 0.9 - 16 | <36 |
| 6 | Chol | Cholesterol | Numeric | mmol/l | 0.0-10.3 | <5 |
| 7 | TG | Triglycerides | Numeric | mmol/l | 0.3 – 13.8 | <2 |
| 8 | HDL | High Density Lipoprotein | Numeric | mmol/l | 0.2 – 9.9 | <1 |
| 9 | LDL | Low Density Lipoprotein | Numeric | mmol/l | 0.3 – 9.9 | <3 |
| 10 | VLDL | Very Low Density Lipoprotein | Numeric | mmol/l | 0.1 - 35 | .1-1.7 |
| 11 | BMI | Body Mass index | Numeric |  | 19 - 47 | 18.5 – 24.9 |
| 12 | Class |  | Character | N (no-diabetes)  P (pre-diabetic)  D (diabetic) | 103 no-diabetes; 53 pre-diabetic; and 844 diabetic | n/a |

Figure 5. Data Information

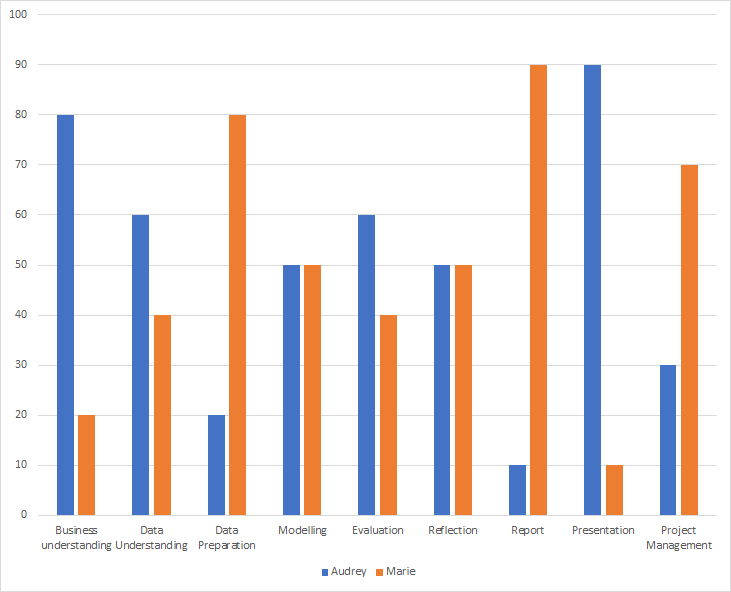
* 1. **Flow Chart of Modeling**



* 1. **Gerate test design**
     1. Split the data into training and testing sets
     2. Select an appropriate algorithm
     3. Train the model on the training data
     4. Evaluate the model's performance on the testing data
  2. **Build a model**
     1. Random Forests was used successfully with an accuracy rate of Accuracy: 0.9966666666666667.
     2. Decision Tree was not used successfully as it kept presenting errors.
  3. **Assess the model**
     1. An accuracy of 0.9966666666666667 is very high. This is good but could indicated overfitting to the training set.

### **Evaluation**

* 1. Based on the analysis of the dataset and the results above, we think that is it possible to use machine learning to predict the association of obesity with diabetes.
  2. A larger dataset would provide more information.
  3. Our team worked well on this project and the distribution of team input is shown in figure 5.

*Figure 5 Team Contributions*

### **Conclusion**

* 1. Having done this initial work, and having learned from it, what we would like to do next is:
     1. Re-run the tests using different training splits (10/20/30%) to see if we can get a more accurate model.
     2. Find a new, larger, more complex dataset and make predictions on this data using the trained model
     3. We could also re-run the tests but change the hyper parameters, i.e. the settings we chose when applying the ML algorithms. For example, we could change the number of trees selected in a random forest, or the kernel type and regularisation parameter in a support vector machine (SVM).

### **References**

Blüher, M. *Obesity: global epidemiology and pathogenesis*. Nat Rev Endocrinol 15, 288–298 (2019). <https://doi.org/10.1038/s41574-019-0176-8>

Cai XT, Ji LW, Liu SS, Wang MR, Heizhati M, Li NF. Derivation and Validation of a Prediction Model for Predicting the 5-Year Incidence of Type 2 Diabetes in Non-Obese Adults: A Population-Based Cohort Study. Diabetes Metab Syndr Obes. 2021;14:2087-2101. <https://doi.org/10.2147/DMSO.S304994>

Draw.io accessed at : https://app.diagrams.net/

Eurostat Overweight and obesity - BMI statistics (2021) Available <https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Overweight_and_obesity__BMI_statistics#Obesity_in_the_EU:_gender_differences>

Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri Ramírez, C.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez Quintana, MJ.... (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering. 33(8):3048-3061. https://doi.org/10.1109/TKDE.2019.296268

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media.

Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1

Rajput, M.R. and Khedgikar, S.S., 2022. Diabetes prediction and analysis using medical attributes: A Machine learning approach. Journal of Xi’an University of Architecture & Technology, 14(1), pp.98-103.

Sarathambekai, S., Vairam, T., Kumar, P., Varun, R. and Jawhar, M.G., 2022, March. A Comparative Analysis of Diabetes Prediction Models using Machine Learning Algorithms. In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 261-265). IEEE.

Safefood, (2012). The cost of overweight and obesity on the island of Ireland. Executive summary. https://www.safefood.net/research-reports/cost-overweight-obesity [Accessed 27 Nov 2021]

Singh, D.A.A.G., Leavline, E.J. and Baig, B.S., 2017. Diabetes prediction using medical data. Journal of Computational Intelligence in Bioinformatics, 10(1), pp.1-8.Tejas N. Joshi & Pramila M. Chawan, Diabetes Prediction Using Machine Learning Techniques, Int. Journal of Engineering Research and Application www.ijera.com ISSN: 2248-9622, Vol. 8, Issue 1, (Part - II) January 2018, pp.-09-13.

Sisodia Deepti and Sisodia Dilip Singh, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, Volume 132, 2018, Pages 1578-1585, ISSN 187 0509, https://doi.org/10.1016/j.procs.2018.05.122

Soni, Mitushi and Sunita Varma. “Diabetes Prediction using Machine Learning Techniques.” International journal of engineering research and technology 9 (2020): n. pag.

Tan C, Li B, Xiao L, Zhang Y, Su Y, Ding N. (2022) A Prediction Model of the Incidence of Type 2 Diabetes in Individuals with Abdominal Obesity: Insights from the General Population. Diabetes Metab Syndr Obes. 2022;15:3555-3564 https://doi.org/10.2147/DMSO.S386687

Thamrin, S.A., Arsyad, D.S., Kuswanto, H., Lawi, A. and Nasir, S., 2021. Predicting obesity in adults using machine Learning techniques: An analysis of Indonesian Basic Health Research 2018. Frontiers in nutrition, 8, p.669155.

Woldaregay, A.Z., Årsand, E., Walderhaug, S., Albers, D., Mamykina, L., Botsis, T. and Hartvigsen, G., 2019. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial intelligence in medicine*, *98*, pp.109-134.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M. and Suez, J., 2015. Personalized nutrition by prediction of glycemic responses. *Cell*, *163*(5), pp.1079-1094.

**Annex A: Data Dictionary**

| **Col** | **Abbreviation** | **Full name of Variable** | **Definition of Variable** |  | **Reported Range** | **Normal Range** | **Variable** | | **Input or derived?** | **Missing Data?** | **Treatment in Preprocessing** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | ID | Study ID number | No assigned during study | number | Any |  | Numeric | Scale | Input | None | Dropped as unique number |
| **B** | No\_Pation | Patient No | No assigned during medical treatment | number | Any |  | Numeric | Scale | Input | None | Dropped as unique number |
| **Col** | Gender | Gender | Gender of subject | 1=Male; 2=Female | M and F | M and F | String | Nominal | Input | None | Recoded to two columns, Age\_Male and Age\_Female |
| **D** | AGE | Age | Age of subject | Years | 20 – 79 | Up to 100 | Numeric | Scale | Input | None | Maintained |
| **E** | Urea | Urea | Clinical measurement indicating functioning of kidneys | mg/dl | 0.5 - 38.9 | 2.5 – 8.5 | Numeric | Scale | Input | None | Maintained |
| **F** | Cr | Creatinine | Clinical measurement indicating functioning of kidneys | micromol/l | 48 – 80 | 49-90 | Numeric | Scale | Input | None | Maintained |
| **G** | HbA1c | Haemoglobin A1c | Clinical measurement used to diagnose diabetes | mmol/l | 0.9 - 16 | <36 | Numeric | Scale | Input | None | Maintained |
| **H** | Chol | Cholesterol | Clinical measurement used to diagnose heart disease | mmol/l | 0.0-10.3 | <5 | Numeric | Scale | Input | None | Maintained |
| **I** | TG | Triglycerides | Clinical measurement used to diagnose heart disease | mmol/l | 0.3 – 13.8 | <2 | Numeric | Scale | Input | None | Maintained |
| **J** | HDL | High Density Lipoprotein | Clinical measurement used to diagnose heart disease | mmol/l | 0.2 – 9.9 | <1 | Numeric | Scale | Input | None | Maintained |
| **K** | LDL | Low Density Lipoprotein | Clinical measurement used to diagnose heart disease | mmol/l | 0.3 – 9.9 | <3 | Numeric | Scale | Input | None | Maintained |
| **L** | VLDL | Very Low Density Lipoprotein | Clinical measurement used to diagnose heart disease | mmol/l | 0.1 - 35 | .1-1.7 | Numeric | Scale | Input | None | Maintained |
| **M** | BMI | Body Mass index | Measure using height and weight to determine if body weight is healthy | number | 19 - 47 | 18.5 – 24.9 | Numeric | Scale | Input | None | Maintained |
| **N** | CLASS | CLASS | Assigned during study | N = no-diabetes; P = pre-diabetic; D = Diabetic | 103 no-diabetes; 53 pre-diabetic; and 844 diabetic | n/a | String | Nominal | Input | None | Recoded to two columns, Diabetic and Not Diabetic |